Family-based quantitative trait meta-analysis implicates rare noncoding variants in *DENND1A* in pathogenesis of polycystic ovary syndrome

Matthew Dapas[1], Ryan Sisk[1], Richard S. Legro[2], Margrit Urbanek[1,3,4], Andrea Dunaif[5,¶*], M. Geoffrey Hayes[1,3,6¶*]

[1] Division of Endocrinology, Metabolism, and Molecular Medicine, Department of Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[2] Department of Obstetrics and Gynecology, Penn State College of Medicine, Hershey, PA

[3] Center for Genetic Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL

[4] Center for Reproductive Science, Northwestern University Feinberg School of Medicine, Chicago, IL

[5] Division of Endocrinology, Diabetes and Bone Disease, Icahn School of Medicine at Mount Sinai, New York, NY

[6] Department of Anthropology, Northwestern University, Evanston, IL

¶ These authors jointly supervised this work.

* Corresponding author

E-mail: ghayes@northwestern.edu (MGH)

andrea.dunaif@mssm.edu (AD)

ABSTRACT

Polycystic ovary syndrome (PCOS) is among the most common endocrine disorders of premenopausal women, affecting 5-15% of this population depending on the diagnostic criteria applied. It is characterized by hyperandrogenism, ovulatory dysfunction and polycystic ovarian morphology. PCOS is a leading risk factor for type 2 diabetes in young women. PCOS is highly heritable, but only a small proportion of this heritability can be accounted for by the common genetic susceptibility variants identified to date. To test the hypothesis that rare genetic variants contribute to PCOS pathogenesis, we performed whole-genome sequencing on DNA from 62 families with one or more daughters with PCOS. We tested for associations of rare variants with PCOS and its concomitant hormonal traits using a quantitative trait meta-analysis. We found rare variants in *DENND1A* (P=5.31×10^{-5}, P_{adj}=0.019) that were significantly associated with reproductive and metabolic traits in PCOS families. Common variants in *DENND1A* have previously been associated with PCOS diagnosis in genome-wide association studies. Subsequent studies indicated that *DENND1A* is an important regulator of human ovarian androgen biosynthesis. Our findings provide additional evidence that *DENND1A* plays a central role in PCOS and suggest that rare noncoding variants contribute to disease pathogenesis.

INTRODUCTION

Polycystic ovary syndrome (PCOS) is a common endocrine disorder affecting 5-15% of premenopausal women worldwide[1], depending on the diagnostic criteria applied. PCOS is diagnosed by two or more of its reproductive features of hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology. It is frequently associated with insulin resistance and pancreatic β-cell dysfunction, making it a leading risk factor for type 2 diabetes in young women[2].

PCOS is a highly heritable complex genetic disorder. Analogous to other complex traits[3], common susceptibility loci identified in genome-wide association studies (GWAS)[4-8] account for only a small proportion of the estimated genetic heritability of PCOS[9]. As GWAS were designed to assess common allelic variants, usually with minor allele frequencies (MAF) ≥2-5%, it has been proposed that less frequently occurring variants with greater effect sizes

account for the observed deficit in heritability[10]. Next generation sequencing approaches have identified rare variants that contribute to complex disease pathogenesis[11-16].

We tested the hypothesis that rare variants contribute to PCOS by conducting family-based association analyses using whole-genome sequencing data. We filtered and weighted rare variants (MAF ≤2%) according to their predicted levels of deleteriousness and grouped them regionally and by genes. We not only tested for associations with PCOS diagnosis, but also with its correlated, quantitative reproductive and metabolic trait levels: testosterone (T), dehydroepiandrosterone sulfate (DHEAS), luteinizing hormone (LH), follicle stimulating hormone (FSH), sex hormone binding globulin (SHBG), and fasting insulin (I). We then combined the quantitative trait phenotypes using a meta-analysis approach to identify sets of rare variants that associate with altered hormonal levels in PCOS.

SUBJECTS AND METHODS

Subjects

This study included 261 individuals from 62 families with PCOS who were Caucasians of European ancestry. Families were ascertained by an index case who fulfilled the National Institutes of Health (NIH) criteria for PCOS[17]. Each family consisted of at least a proband-parent trio. Brothers were not included in this study. Phenotypic data and some genetic analyses on these subjects have been previously reported[18-20]. The study was approved by the Institutional Review Boards of Northwestern University Feinberg School of Medicine, Penn State Health Milton S. Hershey Medical Center, and Brigham and Women's Hospital. Written informed consent was obtained from all subjects prior to study.

Phenotyping for the dichotomous trait analysis (affected vs. unaffected) was performed as previously described[21]. Women were ages 14-63 years, in good health and not taking medications known to alter reproductive or metabolic hormone levels for at least one month prior to study. They had each had both ovaries and a uterus. Exogenous gonadal steroid administration was discontinued at least three months prior to the study. Thyroid, pituitary, and adrenal disorders were excluded by appropriate tests[21]. Women were considered to be of reproductive age if they were between the ages of at least 2 years post-menarche and 45 years old, and had FSH levels ≤40mIU/mL. Hyperandrogenemia was defined by elevated levels of T

(>58 ng/dL), non-SHBG bound T (uT; >15 ng/dL), and/or DHEAS (>2683 ng/mL). Ovarian dysfunction was defined as ≤6 menses per year. Ovarian morphology was not assessed because it does not correlate with the endocrine phenotype[6, 22].

Reproductive-age women with hyperandrogenemia and ovarian dysfunction were assigned a PCOS phenotype. Reproductive-age women with normal androgen levels and regular menses (every 27-35 days) were assigned an unaffected phenotype. Reproductive-age women with hyperandrogenemia and regular menses were assigned a hyperandrogenemic (HA) phenotype. Because androgen levels do not decrease during the menopausal transition[23], women between 46 -63 years with HA were also assigned the HA phenotype, regardless of menstrual cycle pattern. One index case fulfilled the criteria for PCOS when she was 45 years but she was 46 years when enrolled in the study. She was confirmed to have persistent HA and ovarian dysfunction. As done in our previous linkage[22] and family-based association testing[24] studies, women with both PCOS and HA phenotypes were considered affected. In the present study, we also included HA women between 46-63 years as affected. Women with normal androgen levels who were not of reproductive age and all fathers in the study were not assigned a phenotype.

The quantitative trait analysis examined associations between rare variants and T, DHEAS, SHBG, LH, FSH and insulin levels. In addition to the women included in the dichotomous trait analysis, women were included for quantitative trait association testing as follows. No additional women were included in the LH or FSH analyses. Women 46-72 years old were included in the analyses for T and DHEAS since androgen levels do not change during the menopausal transition[23]. These women were also included in the analysis of SHBG and insulin. Women with bilateral oophorectomy (n=10) not receiving postmenopausal hormone therapy were included in the analysis of the adrenal androgen, DHEAS, and insulin[25, 26]. We compared hormonal traits in women receiving postmenopausal hormone therapy (n=15) to women from the cohort of comparable age who were not on receiving hormonal therapy (n=10). Only SHBG levels differed significantly. Therefore, women receiving postmenopausal hormone therapy were included in the analyses of T and DHEAS. They were not included in the insulin analysis because of the effect of estrogen on circulating insulin levels[27].

Fathers were included in the insulin level association test since there are not sex differences in this parameter[28]. Subjects receiving glucocorticoids (men=0, women=2) were excluded from all

quantitative trait association tests. Where applicable, subjects receiving anti-diabetic medications (men=6, women=7) were excluded from the T, SHBG, and insulin analyses but were included in the DHEAS analysis[29]. Subjects with type 2 diabetes not receiving medications (men=3, women=7) were excluded from the insulin analysis but the women were included in the T, DHEAS and SHBG analyses.

Reference ranges for hormonal parameters from concurrently studied reproductively normal control subjects of comparable age, sex, BMI and ancestry[6, 30, 31] are included in **Table 1**. All control subjects had normal glucose tolerance with a 75g oral glucose tolerance test[32]. Reproductive age control women were 18-45 years with regular menses, FSH<40 mIU/ml and normal androgen levels. Older control women were 46-65 years with a history of regular menses and normal androgen levels. Control men were 46- 65 years.

Hormone assays

T, DHEAS, sex hormone binding globulin (SHBG), luteinizing hormone (LH), follicle-stimulating hormone (FSH), and insulin levels were measured as previously reported[6].

Whole-genome sequencing (WGS)

Genomic DNA was isolated from whole blood samples using the Gentra Puregene Blood Kit (Qiagen, Valencia, CA). Whole-genome sequencing was performed by Complete Genomics, Inc., (CGI) using their proprietary sequencing technology. Their sequencing platform employed high-density nanoarrays populated with amplified DNA clusters called DNA nanoballs. DNA was read using a novel, iterative hybridization and ligation approach, which produces paired-end reads up to 70 bases in length[33]. CGI's sequencing service included sample quality control, library construction, whole genome DNA sequencing, and variant calling. Reads were mapped to the NCBI Build 37.2 reference genome (GRCh37).

Variant calling

Variant calling was performed using CGI's Assembly Pipeline version 2.0. Although raw read data were provided, because of the unique gapped read structure produced by CGI sequencing, the use of other mapping or variant-call software was not recommended by CGI. Results were provided in CGI's unique variant-call format. Variants included single nucleotide variants (SNVs), as well as insertions, deletions, and substitutions ≤50 bases in length. Calls were assigned confidence scores assuming equal allele fractions for the diploid genome

(*varScoreEAF*). A Bayesian probability model was used to evaluate potential locus calls. The model accounted for read depth, base call quality scores, mapping/alignment probabilities, and empirical priors on gap sizes and discordance rates[34]. Based on the relative allele likelihoods a quality score was assigned for the chosen call:

$$varScore = 10 \cdot \log_{10}\left(\frac{P(Call\ is\ true)}{P(call\ is\ false)}\right) \quad (1)$$

Scores were therefore reported in decibels (dB), where a score of 10dB represents a likelihood ratio of 10:1, 20dB means 100:1 likelihood, 30dB means 1000:1, etc. Quality thresholds for reporting variants were minimized (≥10dB for homozygous and ≥20dB for heterozygous variant calls) in order to maximize sensitivity. Variants were assigned a basic high/low quality flag (*VQHIGH* or *VQLOW*) based on a quality score threshold of 20dB for homozygous and 40dB for heterozygous variant calls.

Called variant filtering

Variants were considered rare if they appeared with a frequency ≤2% in the Scripps Wellderly Genome Resource, which consisted of 597 unrelated participants of European Ancestry from the Scripps Wellderly Study[35]. The Wellderly study population is composed entirely of elderly individuals ≥80 years of age with no history of chronic disease. Within each sequenced family, reported variants that were inconsistent with Mendelian patterns inheritance were removed from consideration. A variant was considered consistent with Mendelian inheritance if it was called (*VQHIGH*) in one or more of the offspring and in at least one parent. The vast majority of DNA sequencing errors can be eliminated using Mendelian inheritance analysis[36]. As previously described[37-40], an additional set of filters was applied to called variants for each sample in order to minimize the number of false positive calls: (i) variants with *VQLOW* allele tags were removed; (ii) variants in microsatellite regions were removed; (iii) variants within simple tandem repeat regions were removed; (iv) three or more SNPs clustered within a distance of 10bp were removed; (v) SNPs located within 10bp of an insertion or deletion (indel) were removed; (vi) calls located within known regions of segmental duplication were removed; (vii) calls with an observed read depth greater than 3× the average read depth (>168) were removed.

Selection of optimal read depth and quality score thresholds

After systematically applying the filter matrix outlined above, optimal read depth and quality score thresholds were determined for each variant type by comparing calls between replicated samples for a particular family that was sequenced twice by CGI. Variants that were concordant between offspring sample pairs—above a given coverage depth and quality score threshold—were considered as true positive (TP) calls, while those that were discordant between replicates were considered false positives (FP). Accordingly, concordant variant calls that fell below a given depth and quality threshold were classified as false negative (FN) and discordant calls that fell below a given depth and quality threshold were classified as true negative (TN). Optimal depth and quality score thresholds were determined by selecting the thresholds that yielded the greatest Matthews correlation coefficient (MCC) values across each variant type[37]:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}} \qquad (2)$$

The thresholds chosen for association testing corresponded with the greatest MCC values for each variant type (**Table S2**).

Assessing deleteriousness

After filtering for rare variants called with high confidence, as described above, variants were further characterized according to their predicted effects. Only variants that were predicted to have a deleterious effect were included in the association testing. Deleteriousness was primarily assessed using the Combined Annotation Dependent Depletion (CADD) tool[41], which is trained on output from numerous annotation programs to predict deleteriousness based on conservation relative to our ancestral genome. Variants were retained if they produced a CADD score greater than the gene-specific mutation significance cutoff (MSC; 95% confidence interval) suggested by Itan *et al.*[42]. For variants outside of gene regions or in inconsistently annotated genes, the mean MSC (MSC_{μ} = 12.889) was used as the cutoff. In a further effort to reduce false positives, only coding variants classified as at least possibly damaging by the PolyPhen2[43] or SIFT[44] variant effect prediction tools and noncoding variants with LINSIGHT[45] scores ≥0.8 were included in our analysis, as integrating individual methods can improve variant effect prediction[42, 46-48]. CADD and LINSIGHT are both primarily based on evolutionary conservation, and therefore carry certain limitations[49]. However, the applicability of prediction tools based on functional annotations from specific cell types[50, 51] is extremely limited for PCOS

because its pathophysiology involves numerous cell types across multiple organs[52] and annotations for relevant cell types are largely unavailable[53-56].

Genome-wide rare variant association testing

Variants were then grouped for association testing using both gene-based and sliding window approaches. Gene regions included all coding and noncoding DNA from the 3' UTR to 7.5kb upstream of the 5' transcriptional start site (TSS) for all RefSeq annotated genes. The sliding window approach included all noncoding variation contained within sequential windows across the genome using three different window sizes: 10kb windows with no overlap, 25kb windows with 12.5kb overlap, and 100kb windows with 75kb overlap.

In rare variant association testing, genes are often filtered according to a minimum number of rare variants detected per gene[57-60] or a minimum cumulative variant frequency (CVF) per gene[61-64] in order to increase power to detect disease associations. In this study, genes and windows were removed from consideration if they did not harbor deleterious variants in at least 10% of affected subjects[65, 66]. For the gene-based test, in order to reduce the resultant bias towards larger genes, the observed CVFs were adjusted for gene length. Coding and noncoding CVFs were modeled separately using linear regressions against the coding sequence length[67, 68] and the square root of noncoding sequence length[69, 70], respectively. The models also accounted for gene-level GC content. The CVFs observed in affected subjects for each gene were then adjusted accordingly prior to applying the 10% threshold. Adjusting for gene length in this manner reduced the risk for erroneous associations, as genes with longer open reading frames are more likely to be reported falsely as disease-associated[71].

Sequence kernel association test (SKAT) statistics were computed using the methods described in Schaid *et al.*[72], which account for pedigree information in calculating trait associations. Custom variant weights were applied for each variant according to their CADD Phred scores:

$$w_v = \frac{C_v - MSC_g}{max_{v,g \in G}\left(C_v - MSC_g\right)} \qquad (3)$$

where C_v is the CADD Phred score for variant v, g is the gene or window, MSC_G is the mutation significance threshold for g, and G is the set of all variants within the genome. In this way, variants that are more likely to be deleteriousness were weighted more heavily than variants where the functional consequences are less likely to be damaging. For the dichotomous trait

239 analysis, the binary outcome was adjusted for age and body mass index (BMI) using a logistic
240 regression.

241 Six quantitative traits, T, DHEAS, LH, FSH, SHBG, and insulin levels, were tested for association.
242 The trait distributions were each positively skewed. Testing against skewed trait distributions
243 can result in heavily inflated Type I error rates in rare variant association testing[73]. Therefore,
244 each trait was modeled against a gamma distribution when adjusting for age and BMI, and non-
245 normal residuals (Shapiro-Wilk < 0.05) were then further normalized using a rank-based
246 inverse normal transformation (INT). The INT has been found to be the optimal method for
247 maintaining Type I error control without sacrificing power in rare variant association testing
248 on non-normally distributed traits[74].

249 **Quantitative trait meta-analysis**

250 For complex diseases with multivariate phenotypes, combining multiple related phenotypic
251 traits into one analysis can increase power in finding disease associations[75, 76]. We combined
252 the six aforementioned quantitative trait associations into one meta-statistic using a Fisher
253 combination function modified to account for correlated traits[77]. Inter-trait correlations were
254 determined using the Pearson correlation coefficient (**Fig. S4**). P-values were adjusted using
255 Bonferroni correction according to the number of variant groupings that were tested that
256 contained at least one variant, as well as by the genomic inflation factor, λ. Correlation between
257 meta-analysis association results and dichotomous trait association results were calculated
258 using Spearman's coefficient.

259 The characteristic disturbance of gonadotropin secretion associated with PCOS is increased LH
260 relative to FSH release[78]. For genes with significant meta-analysis associations, LH:FSH ratios
261 were compared between variant carriers and non-carriers using a Wilcoxon's rank sum test,
262 adjusted for multiple testing (Bonferroni). Differences in LH:FSH ratios between variant
263 carriers and non-carriers would indicate that the gene variants alter gonadotropin signaling.

264 ***In silico* binding effect prediction**

265 To assess the potential functional effects of noncoding variants identified in the quantitative
266 trait meta-analysis, we predicted the corresponding impacts to transcription factor (TF) and
267 RNA-binding protein (RBP) binding *in silico*. For each noncoding SNV identified in the
268 quantitative trait meta-analysis, transcription factor (TF) binding affinities were calculated for

all subsequences overlapping the SNV position on each strand within a ±20bp window. Binding affinity scores were calculated using position weight matrices (PWMs) derived from ENCODE ChIP-Seq experiments[71]. Scores were determined by summing the logged frequencies for a given sequence across a motif PWM. Binding p-values were defined as the probability that a sequence sampled from a genomic background distribution had an affinity score greater than or equal the largest affinity score produced from one of the tested subsequences. Genomic background sequences were generated using a first order Markov model[79]. The significance of a given change in binding affinity scores between reference and SNV alleles was assessed by determining whether the differences in relative binding affinity rank between the two alleles was significantly different than what would be expected by chance[80, 81]. P-values were conservatively adjusted to account for multiple testing using the Benjamini-Hochberg (BH) procedure[82].

Once binding affinities were calculated for each TF at each SNV, filters were applied to identify the most likely candidates for TF binding site disruption. Instances in which the predicted TF binding affinity score was <80% of the maximum affinity score for the given motif were excluded. SNVs in which the reference allele was not predicted to bind a particular TF with statistical significance ($P_{BH}<0.05$) were also removed from consideration, as well as variants in which both the reference and SNV alleles were predicted to bind a TF with statistical significance. Only TFs expressed in the ovary were analyzed. Tissue-specific gene expression was determined using GTEx data[83] (median Reads Per Kilobase of transcript per Million mapped reads [RPKM] ≥0.1).

The identified SNVs were likewise analyzed for potential alteration to RNA-binding protein (RBP) sites following the same procedure but for a few modifications. RBPs and their binding affinity scores were determined using the ATtTRACT database[84]. Only sequences on the coding strand were evaluated as to reflect the mRNA sequences. Additionally, significant changes in RBP binding were considered regardless of the direction of effect, such that instances in which the alternate allele was predicted to induce RBP binding were also included.

Supplementary Methods

For additional details regarding methodological considerations and rationale, please refer to the Appendix.

RESULTS

Characteristics of study population

The characteristics of the study population, including counts and trait distributions by familial relation and the numbers of subjects included in each association test, are summarized in **Table 1**.

Whole-genome sequencing and variant calling

Genome sequencing yielded average genome coverage of 96.2% per sample and an overall mean sequencing depth of 56× (**Fig. S1**). On average, 90.4% and 66.4% of the genome, including 95.2% and 78.8% of the exome, was covered with at least 20× and 40× sequencing depth, respectively. Approximately 4.04 million high-confidence small variant calls were reported per genome. By applying optimal read depth and quality thresholds as well as a series of genomic filters[37], we reduced the discrepancy rate between replicate samples from 0.23% to 0.04% for rare variants (**Tables S1 and S2**).

Association testing and quantitative trait meta-analysis

We found 339 genes that had rare, deleterious variants in at least 10% of cases after adjusting for gene length and %GC content. No set of rare variants reached genome-wide significance for association with PCOS/HA disease status in the dichotomous trait analysis. We found 32 rare variants (2 coding, 30 noncoding) in the *DENND1A* gene that were collectively significantly associated with quantitative trait levels (P=5.31×10^{-5}, P_{adj}=0.019; **Table 2**), after adjusting for multiple testing and for observed genomic inflation (**Fig. S2**). Women with one or more of these *DENND1A* variants had significantly higher LH:FSH ratios (P=0.0012). PCOS/HA phenotype women with one or more *DENND1A* variants had significantly higher LH:FSH ratios than PCOS/HA phenotype women without *DENND1A* variants (P=0.0060). Unaffected women with one or more *DENND1A* variants had higher LH:FSH ratios than unaffected women without *DENND1A* variants (P=0.0586; **Fig. 1**), but the difference was not

significant after multiple test correction (P<0.0167). No other gene-based set of rare variants reached genome-wide significance for association with quantitative trait levels. The correlation between the meta-analysis gene associations and the dichotomous trait gene associations was 0.24 (Spearman).

Using the sliding windows approach, we found a subset of noncoding variants within a 25kb region of *DENND1A* (chr9:126,537,500-126,562,500) that were significantly associated with altered quantitative trait levels (P=1.92×10^{-5}, P_{adj}=9.53×10^{-3}; **Fig. S3**; **Supplementary Data**). The region included three noncoding variants that were collectively present in eight PCOS/HA subjects and zero unaffected subjects. One of these variants, rs117893097 ($MAF_{Wellderly}$ = 0.013), was homozygous in one of the subjects. This 25kb region encompasses one of the *DENND1A* GWAS risk variants (rs10986105; $MAF_{Wellderly}$ = 0.034; OR_{Meta}=1.39[85]), although none of the subjects with one of the rare variants in the region also had the rs10986105 risk allele. The relative positions of all of the rare variants found in *DENND1A* are shown in **Fig. 2**. No other windows across the genome were found to have significant associations with disease state or hormonal levels.

Several other PCOS GWAS candidate genes appeared in our filtered set of genes, including *C9orf3*[5, 6, 8] (P=6.14×10^{-3}), *HMGA2*[5] (P=0.062), *ZBTB16*[8] (P=0.20), *TOX3*[5, 8] (P=0.22), and *THADA*[4, 5, 7, 8] (P=0.74). *C9orf3* had the 4th strongest association overall (**Supplementary Data**), but failed to reach genome-wide significance after correction for multiple testing. The relative quantitative trait associations for these genes are illustrated in **Fig. 3**.

Protein binding effect prediction

Nine of the *DENND1A* variants were predicted to significantly impact TF binding motifs, while the majority of variants were predicted to significantly alter RBP binding motifs (17 disrupted, 14 induced). The specific TF and RBP motifs associated with each noncoding variant are listed in **Table 3**. Binding by the heterogeneous nuclear ribonucleoprotein (hnRNP) family of RBPs appeared to be the most commonly impacted.

DISCUSSION

We identified rare variants in *DENND1A* that were significantly associated with altered reproductive and metabolic hormone levels in PCOS. These findings are of considerable

356 interest because common SNVs in *DENND1A* were associated with PCOS diagnosis in a GWAS of
357 Han Chinese women[4]; these associations were subsequently replicated in women of European
358 ancestry[86]. Our study, using an independent family-based WGS analytical approach, provides
359 further evidence that *DENND1A* is an important gene in the pathogenesis of PCOS. These
360 findings complement the studies of McAllister and colleagues[56, 87] that have shown that
361 *DENND1A* plays a key role in androgen biosynthesis in human theca cells and is upregulated in
362 PCOS theca cells.

363 *DENND1A* encodes a protein that is a member of the connecdenn family of proteins, which
364 function as guanine nucleotide exchange factors for the Rab family of small GTPases[77]. The
365 DENND1A protein, also known as Connecdenn 1, is thought to link Rab35 activation with
366 clathrin-mediated endocytosis[42]. Following its reported associations with PCOS[4, 86], McAllister
367 and colleagues[87] investigated the role of *DENND1A* in ovarian androgen biosynthesis, a key
368 biologic pathway that is disrupted in PCOS[88]. *DENND1A* encodes two transcripts as the result of
369 alternative splicing, DENND1A.V1 and DENND1A.V2[56]. The encoded V2 protein was found in
370 ovarian theca cells and its abundance was correlated with increased androgen production[87].
371 The expression of V2 was increased in PCOS theca cells[87]. Forced expression of the *V2*
372 transcript produced a PCOS phenotype in normal theca cells, whereas knockdown of *V2* in
373 PCOS theca cells reduced thecal androgen biosynthesis[87]. Urine exosomal V2 mRNA was also
374 increased in PCOS women[87]. Taken together, these findings provide strong support for the
375 hypothesis that *DENND1A* plays a role in PCOS pathogenesis. The increased LH:FSH ratios that
376 we observed in the *DENND1A* rare variant carriers (**Fig. 1**) suggest that *DENND1A* plays a role
377 in the regulation of gonadotropin secretion[89].

378 The *DENND1A* risk variants identified by GWAS are located in introns and the functional
379 consequences of these variants are unknown[56]. There have been no large-scale sequencing
380 studies reported to map causal variants in *DENND1A*[85]. Targeted[87, 90] and whole exome
381 sequencing[91] in small cohorts of PCOS women have failed to identify any coding variants in
382 *DENND1A* that were associated with PCOS or with V2 isoform expression. Genomic sequencing
383 of the intronic region where V1 and V2 are alternatively spliced also failed to identify any
384 variants that consistently favored V2 expression in a study of 20 normal and 19 PCOS women[56].
385 The GWAS risk variants and most of the variants identified in our study lie well upstream (100-
386 400kb) of this region (**Fig. 2**).

Many of the *DENND1A* variants identified in the present study were predicted to disrupt conserved TF binding motifs, which could affect gene expression, but most of the variants were predicted to alter affinities of RBPs to the mRNA transcript (**Table 3**). It is plausible, therefore, that the rare variants we reported were selectively driving the expression of the V2 splicing variant via post-transcriptional regulation. Collectively, the *DENND1A* variants identified in this study were found in 50% of families, but each individual variant was typically found in only one or two families. Our findings, therefore, support a model of PCOS in which causal variants are individually uncommon but collectively tend to occur in key genes. Our recent findings[92] of multiple rare exonic rare variants in the *AMH* gene that reduce its biologic activity in ~3% of women with PCOS is consistent with this model.

Our results also align with the emerging evidence that rare coding variants with large effect sizes do not play a major role in complex disease[93]. Rather, it appears that complex traits are primarily driven by noncoding variation[94, 95], both common and rare[96, 97]. Of the 32 rare variants predicted as deleterious that we identified in *DENND1A*, 30 were noncoding. Rare variant association studies typically require very large sample sizes[98], but paired WGS and transcriptome sequencing analysis from one large family demonstrated that rare noncoding variants have strong effects on individual gene-expression profiles[99]. In a similarly designed study to ours, Ament and colleagues[100] identified rare variants associated with increased risk of bipolar disorder, the vast majority of which were noncoding. Due to the relative cost-effectiveness of whole exome sequencing[94], the limited availability of computational tools designed to predict the effects of noncoding variants on phenotypes[101], and our relatively poor understanding of regulatory mechanisms in the genome[102], noncoding variants have been noticeably understudied in complex trait genetics. As larger WGS datasets are accumulated and the focus of complex trait studies shifts more towards understanding regulatory mechanisms, the contribution of rare noncoding variants in various complex diseases will become clearer.

Several established PCOS candidate genes besides *DENND1A* appeared among the top gene associations, but failed to reach genome-wide significance. These genes included previously reported PCOS GWAS susceptibility loci, *C9orf3*, *HMGA2*, *ZBTB16*, *TOX3, and THADA* (**Fig. 3**). Two additional genes with strong, but not genome-wide-significant, associations with PCOS quantitative traits are highly plausible PCOS candidate genes. *BMP6* had the third strongest association in our meta-analysis (P=4.00×10^{-3}, **Table 2**). It is a member of the bone

morphogenetic protein family, which are growth factors involved in folliculogenesis. *BMP6* expression was previously found to be significantly higher in granulosa cells from PCOS women compared with reproductively normal control women[103]. Moreover, BMP6 was found to increase expression of the FSH receptor, inhibin/activin β subunits, and AMH genes in human granulosa cells[104]. *PRDM2* had the fifth strongest association in our meta-analysis (P=6.95×10^{-3}, **Table 2**). *PRDM2* is an estrogen receptor co-activator[105] that is highly expressed in the ovary and pituitary gland[83]. Ligand bound estrogen receptor alpha (ERα) binds with PRDM2 to open chromatin at ERα target genes[105, 106]. PRDM2 also binds with the retinoblastoma protein[107], which has been shown to play an important role in follicular development in granulosa cells[108, 109].

The central statistical challenge in studying rare variants is achieving adequate power to detect significant associations while controlling for Type I error. The family-based structure of our cohort provided an enrichment of individual rare variants and enabled modeling of familial segregation[110]. To mitigate variant calling errors, we utilized replicate samples from one family to determine optimal read depth and quality thresholds. To remove irrelevant variants from consideration, we applied a LINSIGHT score threshold and gene-specific CADD score thresholds[72] to filter for deleteriousness. We further prioritized variants by weighting them by their relative CADD scores. To group rare variants effectively, we applied several windows-based binning methods, in addition to the gene-based approach, to ensure that different kinds of functionally-correlated genomic regions were tested, both of fixed length and of variable length. To limit our search to genes that were more likely to have specific roles in PCOS etiology, we only considered genes with rare deleterious variants in at least 10% of cases, as causal rare variants are more likely to accumulate in core disease genes[94]. We greatly increased our power to detect relevant disease genes and account for pleiotropic effects by consolidating quantitative trait association results into a meta-analysis[75, 76]. In order to reduce Type I error, in addition to the variant calling quality control measures, we modeled the quantitative traits against skewed distributions and further normalized trait residuals using an INT.

Given the size of our cohort, it was necessary to apply relatively strict filters based on predicted variant effects and cumulative allele frequencies in order to detect rare variant associations. Very large sample sizes are otherwise required for WGS studies of rare variants[98]. By only including genes with rare, likely-deleterious variants in at least 10% of cases, we greatly

reduced the multiple-testing burden of our analysis, thereby increasing our power to detect core PCOS genes[111, 112]. Applying *a priori* hypotheses regarding which variants and genes may be relevant to disease, however, is analogous to a candidate gene approach[98]. Any sets of rare variants that contribute to PCOS in smaller subpopulations of PCOS women, as we found for *AMH*[92], were likely removed from consideration. Likewise, any causal rare variants that were not predicted bioinformatically to be deleterious based on existing annotations and evolutionary conservation[41, 45] would not have been detected. Furthermore, because variants were filtered for consistency with Mendelian inheritance, *de novo* mutations were not considered.

Because the number of genetic variants is directly correlated with the size of the gene[71], the CVF threshold introduced a bias towards larger genes in some of the analyses. However, this bias was mitigated by adjusting for gene length. Furthermore, any such bias was not applicable to our fixed-length windows-based approach, which replicated our *DENND1A* findings. It is also possible that causal rare variants with larger effect sizes were omitted from the meta-analysis because we tested against normalized trait residuals in an effort to reduce Type I errors. Using normalized trait residuals may have excluded variants with large effects that produced outliers. However, as mentioned above, recent evidence has demonstrated that complex traits are primarily driven by noncoding variation with modest effect sizes[94, 95].

Despite the numerous steps taken to increase power, our study ultimately remained underpowered to detect rare variant associations with PCOS diagnosis in the dichotomous trait analysis. Although an association with PCOS quantitative traits implicates genetic variants in disease pathogenesis, it does not necessarily mean that a gene is associated with PCOS itself. The correlation between the dichotomous trait results and quantitative trait meta-analysis results was 0.24. The noncoding variants identified in this study, despite being rare and predicted to be deleterious, may be in linkage disequilibrium with the actual pathogenic variant on the same alleles that were removed by the applied allele frequency or predicted effect thresholds. Replication and functional studies are needed to confirm individual variant functionality and disease associations.

In summary, by applying family-based sequence kernel association tests on filtered whole-genome variant call data from a cohort of PCOS families, we were able to identify rare variants in the *DENND1A* gene that were associated with quantitative hormonal traits of PCOS. Our

480 results suggest that rare noncoding variants contribute to the distinctive hormonal profile of
481 PCOS. This study also demonstrates that using a quantitative trait meta-analysis can be a
482 powerful approach in rare variant association testing, particularly for complex diseases with
483 pleiotropic etiologies.

ACKNOWLEDGEMENTS

This study was supported by US National Institutes of Health (NIH) grants P50 HD044405 (A.D.) and R01 HD085227 (A.D.). M.D. was supported by NRSA fellowship T32 DK007169. This study uses data from the Scripps Wellderly Genome Resource, which is funded under NIH grant 5 UL1 TR001114 Scripps Translational Science Institute CTSA Award. We thank Drs. Terry Farrah and Gustavo Glusman, from the Institute for Systems Biology for facilitating variant queries via Kaviar (http://db.systemsbiology.net/kaviar). We thank the Northwestern University Research Computing Services team for supporting the computational needs of this research. We are very grateful to all of the families who took part in the study.

AUTHOR CONTRIBUTIONS

M.D. M.U., A.D, and M.G.H. conceived and designed the study. M.D. performed experiments and statistical analyses. R.S. aided with data management, project coordination, and statistical analyses. R.S.L. and A.D. recruited study subjects and measured or analyzed phenotypic data. M.D. wrote the paper. All authors critically reviewed and approved the paper.

REFERENCES

1. Azziz, R. PCOS in 2015: New insights into the genetics of polycystic ovary syndrome. *Nat Rev Endocrinol* **12**, 74-75 (2016).
2. Gambineri, A. *et al.* Polycystic ovary syndrome is a risk factor for type 2 diabetes: results from a long-term prospective study. *Diabetes* **61**, 2369-2374 (2012).
3. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *American journal of human genetics* **99**, 139-153 (2016).
4. Chen, Z.J. *et al.* Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3. *Nature genetics* **43**, 55-59 (2011).
5. Shi, Y. *et al.* Genome-wide association study identifies eight new risk loci for polycystic ovary syndrome. *Nature genetics* **44**, 1020-1025 (2012).
6. Hayes, M.G. *et al.* Genome-wide association of polycystic ovary syndrome implicates alterations in gonadotropin secretion in European ancestry populations. *Nature communications* **6**, 7502 (2015).

7. Day, F.R. *et al.* Causal mechanisms and balancing selection inferred from genetic associations with polycystic ovary syndrome. *Nature communications* **6**, 8464 (2015).
8. Day, F. *et al.* Large-Scale Genome-Wide Meta Analysis of Polycystic Ovary Syndrome Suggests Shared Genetic Architecture for Different Diagnosis Criteria. *bioRxiv* (2018).
9. Vink, J.M., Sadrzadeh, S., Lambalk, C.B. & Boomsma, D.I. Heritability of polycystic ovary syndrome in a Dutch twin-family study. *The Journal of clinical endocrinology and metabolism* **91**, 2100-2104 (2006).
10. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).
11. Li, X. *et al.* The impact of rare variation on gene expression across tissues. *Nature* **550**, 239-243 (2017).
12. Consortium, U.K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82-90 (2015).
13. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nature genetics* **47**, 589-597 (2015).
14. Zou, Y. *et al.* IRX3 Promotes the Browning of White Adipocytes and Its Rare Variants are Associated with Human Obesity Risk. *EBioMedicine* **24**, 64-75 (2017).
15. Turcot, V. *et al.* Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nature genetics* **50**, 26-41 (2018).
16. Russo, A., Di Gaetano, C., Cugliari, G. & Matullo, G. Advances in the Genetics of Hypertension: The Effect of Rare Variants. *Int J Mol Sci* **19** (2018).
17. Zawadski, J.K.D., A. in Polycystic Ovary Syndrome 377-384 (Blackwell Scientific, Boston, Massachussets; 1992).
18. Diamanti-Kandarakis, E. & Dunaif, A. Insulin resistance and the polycystic ovary syndrome revisited: an update on mechanisms and implications. *Endocr Rev* **33**, 981-1030 (2012).
19. Kosova, G. & Urbanek, M. Genetics of the polycystic ovary syndrome. *Mol Cell Endocrinol* **373**, 29-38 (2013).
20. Dunaif, A. Perspectives in Polycystic Ovary Syndrome: From Hair to Eternity. *The Journal of clinical endocrinology and metabolism* **101**, 759-768 (2016).
21. Legro, R.S. *et al.* Phenotype and genotype in polycystic ovary syndrome. *Recent Prog Horm Res* **53**, 217-256 (1998).

22. Urbanek, M. *et al.* Thirty-seven candidate genes for polycystic ovary syndrome: strongest evidence for linkage is with follistatin. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 8573-8578 (1999).
23. Kim, C., Harlow, S.D., Zheng, H., McConnell, D.S. & Randolph, J.F., Jr. Changes in androstenedione, dehydroepiandrosterone, testosterone, estradiol, and estrone over the menopausal transition. *Womens Midlife Health* **3** (2017).
24. Urbanek, M. *et al.* Candidate gene region for polycystic ovary syndrome on chromosome 19p13.2. *The Journal of clinical endocrinology and metabolism* **90**, 6623-6629 (2005).
25. Kotsopoulos, J. *et al.* The relationship between bilateral oophorectomy and plasma hormone levels in postmenopausal women. *Horm Cancer* **6**, 54-63 (2015).
26. Yazdani, S., Sharbatdaran, M., Abedi Samakoosh, M., Bouzari, Z. & Masoudi, Z. Glucose Tolerance and lipid profile changes after surgical menopause. *Caspian J Intern Med* **5**, 114-117 (2014).
27. Espeland, M.A. *et al.* Effect of postmenopausal hormone therapy on glucose and insulin concentrations. PEPI Investigators. Postmenopausal Estrogen/Progestin Interventions. *Diabetes Care* **21**, 1589-1595 (1998).
28. Magkos, F., Wang, X. & Mittendorfer, B. Metabolic actions of insulin in men and women. *Nutrition* **26**, 686-693 (2010).
29. Sohrevardi, S.M. *et al.* Evaluating the effect of insulin sensitizers metformin and pioglitazone alone and in combination on women with polycystic ovary syndrome: An RCT. *Int J Reprod Biomed (Yazd)* **14**, 743-754 (2016).
30. Torchen, L.C. *et al.* Increased antimullerian hormone levels and other reproductive endocrine changes in adult male relatives of women with polycystic ovary syndrome. *Fertil Steril* (2016).
31. Sam, S., Legro, R.S., Essah, P.A., Apridonidze, T. & Dunaif, A. Evidence for metabolic and reproductive phenotypes in mothers of women with polycystic ovary syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 7030-7035 (2006).
32. Organization, W.H. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. (2006).
33. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78-81 (2010).
34. Baccash, J., Halpern, A., Tian, C., Pant, K. & Carnevali, P. (Complete Genomics, Inc., US; 2013).
35. Erikson, Galina A. *et al.* Whole-Genome Sequencing of a Healthy Aging Cohort. *Cell*.

36. Roach, J.C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636-639 (2010).
37. Reumers, J. *et al.* Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature biotechnology* **30**, 61-68 (2012).
38. Yuen, R.K. *et al.* Whole-genome sequencing of quartet families with autism spectrum disorder. *Nature medicine* **21**, 185-191 (2015).
39. Genomes Project, C. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
40. Jiang, Y.H. *et al.* Detection of clinically relevant genetic variants in autism spectrum disorder by whole-genome sequencing. *American journal of human genetics* **93**, 249-263 (2013).
41. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310-315 (2014).
42. Itan, Y. *et al.* The mutation significance cutoff: gene-level thresholds for variant predictions. *Nature methods* **13**, 109-110 (2016).
43. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249 (2010).
44. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073-1081 (2009).
45. Huang, Y.F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics* **49**, 618-624 (2017).
46. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137 (2015).
47. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC genomics* **14 Suppl 3**, S7 (2013).
48. Gonzalez-Perez, A. & Lopez-Bigas, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *American journal of human genetics* **88**, 440-449 (2011).
49. Nishizaki, S.S. & Boyle, A.P. Mining the Unknown: Assigning Function to Noncoding Single Nucleotide Polymorphisms. *Trends in genetics : TIG* **33**, 34-45 (2017).
50. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nature protocols* **12**, 2478-2492 (2017).

51. Zhang, Y. & Hardison, R.C. Accurate and reproducible functional maps in 127 human cell types via 2D genome segmentation. *Nucleic acids research* **45**, 9823-9836 (2017).
52. Rosenfield, R.L. & Ehrmann, D.A. The Pathogenesis of Polycystic Ovary Syndrome (PCOS): The Hypothesis of PCOS as Functional Ovarian Hyperandrogenism Revisited. *Endocr Rev* **37**, 467-520 (2016).
53. Consortium, E.P. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
54. Forgacs, A.L. *et al.* BLTK1 murine Leydig cells: a novel steroidogenic model for evaluating the effects of reproductive and developmental toxicants. *Toxicol Sci* **127**, 391-402 (2012).
55. Huang-Doran, I. & Franks, S. Genetic Rodent Models of Obesity-Associated Ovarian Dysfunction and Subfertility: Insights into Polycystic Ovary Syndrome. *Front Endocrinol (Lausanne)* **7**, 53 (2016).
56. Tee, M.K. *et al.* Alternative splicing of DENND1A, a PCOS candidate gene, generates variant 2. *Mol Cell Endocrinol* **434**, 25-35 (2016).
57. Erikson, G.A. *et al.* Whole-Genome Sequencing of a Healthy Aging Cohort. *Cell* **165**, 1002-1011 (2016).
58. Tombacz, D. *et al.* High-Coverage Whole-Exome Sequencing Identifies Candidate Genes for Suicide in Victims with Major Depressive Disorder. *Sci Rep* **7**, 7106 (2017).
59. Wong, J.K.L. *et al.* Rare variants and de novo variants in mesial temporal lobe epilepsy with hippocampal sclerosis. *Neurol Genet* **4**, e245 (2018).
60. He, Z. *et al.* The Rare-Variant Generalized Disequilibrium Test for Association Analysis of Nuclear and Extended Pedigrees with Application to Alzheimer Disease WGS Data. *American journal of human genetics* **100**, 193-204 (2017).
61. Igartua, C. *et al.* Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. *Nature communications* **6**, 5965 (2015).
62. Lubitz, S.A. *et al.* Whole Exome Sequencing in Atrial Fibrillation. *PLoS genetics* **12**, e1006284 (2016).
63. Li, M. *et al.* SOS2 and ACP1 Loci Identified through Large-Scale Exome Chip Analysis Regulate Kidney Development and Function. *J Am Soc Nephrol* **28**, 981-994 (2017).
64. Tg *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *The New England journal of medicine* **371**, 22-31 (2014).

65. Dai, W. *et al.* Whole-exome sequencing identifies MST1R as a genetic susceptibility gene in nasopharyngeal carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **113**, 3317-3322 (2016).
66. Ferre-Fernandez, J.J. *et al.* Whole-Exome Sequencing of Congenital Glaucoma Patients Reveals Hypermorphic Variants in GPATCH3, a New Gene Involved in Ocular and Craniofacial Development. *Sci Rep* **7**, 46175 (2017).
67. Strom, S.P. & Gorin, M.B. Evaluation of autosomal dominant retinal dystrophy genes in an unaffected cohort suggests rare or private missense variants may often be benign. *Mol Vis* **19**, 980-985 (2013).
68. Schroeder, J.W., Hirst, W.G., Szewczyk, G.A. & Simmons, L.A. The Effect of Local Sequence Context on Mutational Bias of Genes Encoded on the Leading and Lagging Strands. *Curr Biol* **26**, 692-697 (2016).
69. Iossifov, I. *et al.* Low load for disruptive mutations in autism genes and their biased transmission. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E5600-5607 (2015).
70. Gao, L., Fang, Z., Zhang, K., Zhi, D. & Cui, X. Length bias correction for RNA-seq data in gene set analyses. *Bioinformatics* **27**, 662-669 (2011).
71. Shyr, C. *et al.* FLAGS, frequently mutated genes in public exomes. *BMC medical genomics* **7**, 64 (2014).
72. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P. & Thibodeau, S.N. Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. *Genet Epidemiol* **37**, 409-418 (2013).
73. Wei, P. *et al.* On Robust Association Testing for Quantitative Traits and Rare Variants. *G3 (Bethesda)* **6**, 3941-3950 (2016).
74. Auer, P.L., Reiner, A.P. & Leal, S.M. The effect of phenotypic outliers and non-normality on rare-variant association testing. *Eur J Hum Genet* **24**, 1188-1194 (2016).
75. Solovieff, N., Cotsapas, C., Lee, P.H., Purcell, S.M. & Smoller, J.W. Pleiotropy in complex traits: challenges and strategies. *Nature reviews. Genetics* **14**, 483-495 (2013).
76. Wu, B. & Pankow, J.S. Fast and Accurate Genome-Wide Association Test of Multiple Quantitative Traits. *Comput Math Methods Med* **2018**, 2564531 (2018).
77. Yang, J.J., Li, J., Williams, L.K. & Buu, A. An efficient genome-wide association test for multivariate phenotypes based on the Fisher combination function. *BMC bioinformatics* **17**, 19 (2016).

78. Lal, L., Bharti, A. & Perween, A. To Study The Status of LH: FSH Ratio in Obese And Non-Obese Patients of Polycystic Ovarian Syndrome.
79. Chan, H.P., Zhang, N.R. & Chen, L.H. Importance sampling of word patterns in DNA and protein sequences. *J Comput Biol* **17**, 1697-1709 (2010).
80. Macintyre, G., Bailey, J., Haviv, I. & Kowalczyk, A. is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics* **26**, i524-530 (2010).
81. Zuo, C., Shin, S. & Keles, S. atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics* **31**, 3353-3355 (2015).
82. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289-300 (1995).
83. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nature genetics* **45**, 580-585 (2013).
84. Giudice, G., Sanchez-Cabo, F., Torroja, C. & Lara-Pezzi, E. ATtRACT-a database of RNA-binding proteins and associated motifs. *Database (Oxford)* **2016** (2016).
85. Gao, J., Xue, J.D., Li, Z.C., Zhou, L. & Chen, C. The association of DENND1A gene polymorphisms and polycystic ovary syndrome risk: a systematic review and meta-analysis. *Arch Gynecol Obstet* **294**, 1073-1080 (2016).
86. Welt, C.K. *et al.* Variants in DENND1A are associated with polycystic ovary syndrome in women of European ancestry. *The Journal of clinical endocrinology and metabolism* **97**, E1342-1347 (2012).
87. McAllister, J.M. *et al.* Overexpression of a DENND1A isoform produces a polycystic ovary syndrome theca phenotype. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E1519-1527 (2014).
88. Legro, R.S., Driscoll, D., Strauss, J.F., 3rd, Fox, J. & Dunaif, A. Evidence for a genetic basis for hyperandrogenemia in polycystic ovary syndrome. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14956-14960 (1998).
89. McCartney, C.R., Eagleson, C.A. & Marshall, J.C. Regulation of gonadotropin secretion: implications for polycystic ovary syndrome. *Semin Reprod Med* **20**, 317-326 (2002).
90. Eriksen, M.B. *et al.* Genetic alterations within the DENND1A gene in patients with polycystic ovary syndrome (PCOS). *PloS one* **8**, e77186 (2013).
91. Khan, M.J., Nazli, R., Ahmed, J. & Basit, S. Whole Genome Sequencing instead of Whole Exome Sequencing is required to identify the Genetic Causes of Polycystic Ovary Syndrome in Pakistani families. *Pak J Med Sci* **34**, 540-545 (2018).

92. Gorsic, L.K. *et al.* Pathogenic Anti-Mullerian Hormone Variants in Polycystic Ovary Syndrome. *The Journal of clinical endocrinology and metabolism* **102**, 2862-2872 (2017).
93. Wray, N.R., Wijmenga, C., Sullivan, P.F., Yang, J. & Visscher, P.M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573-1580 (2018).
94. Boyle, E.A., Li, Y.I. & Pritchard, J.K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186 (2017).
95. Ma, M. *et al.* Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC genomics* **16 Suppl 8**, S3 (2015).
96. Igartua, C., Mozaffari, S.V., Nicolae, D.L. & Ober, C. Rare non-coding variants are associated with plasma lipid traits in a founder population. *Sci Rep* **7**, 16415 (2017).
97. Zhao, J. *et al.* A Burden of Rare Variants Associated with Extremes of Gene Expression in Human Peripheral Blood. *American journal of human genetics* **98**, 299-309 (2016).
98. Wray, N.R. & Gratten, J. Sizing up whole-genome sequencing studies of common diseases. *Nature genetics* **50**, 635-637 (2018).
99. Li, X. *et al.* Transcriptome sequencing of a large human family identifies the impact of rare noncoding variants. *American journal of human genetics* **95**, 245-256 (2014).
100. Ament, S.A. *et al.* Rare variants in neuronal excitability genes influence risk for bipolar disorder. *Proceedings of the National Academy of Sciences of the United States of America* **112**, 3576-3581 (2015).
101. Liu, Q. *et al.* VariFunNet, an integrated multiscale modeling framework to study the effects of rare non-coding variants in Genome-Wide Association Studies: applied to Alzheimer's Disease. *Proceedings (IEEE Int Conf Bioinformatics Biomed)* **2017**, 2177-2182 (2017).
102. Zhou, L. & Zhao, F. Prioritization and functional assessment of noncoding variants associated with complex diseases. *Genome medicine* **10**, 53 (2018).
103. Khalaf, M. *et al.* BMP system expression in GCs from polycystic ovary syndrome women and the in vitro effects of BMP4, BMP6, and BMP7 on GC steroidogenesis. *Eur J Endocrinol* **168**, 437-444 (2013).
104. Shi, J. *et al.* Bone morphogenetic protein-6 stimulates gene expression of follicle-stimulating hormone receptor, inhibin/activin beta subunits, and anti-Mullerian hormone in human granulosa cells. *Fertil Steril* **92**, 1794-1798 (2009).
105. Di Zazzo, E., De Rosa, C., Abbondanza, C. & Moncharmont, B. PRDM Proteins: Molecular Mechanisms in Signal Transduction and Transcriptional Regulation. *Biology (Basel)* **2**, 107-141 (2013).

106. Carling, T. *et al.* A histone methyltransferase is required for maximal response to female sex hormones. *Molecular and cellular biology* **24**, 7032-7042 (2004).
107. Liu, L., Shao, G., Steele-Perkins, G. & Huang, S. The retinoblastoma interacting zinc finger gene RIZ produces a PR domain-lacking product through an internal promoter. *The Journal of biological chemistry* **272**, 2984-2991 (1997).
108. Andreu-Vieyra, C., Chen, R. & Matzuk, M.M. Conditional deletion of the retinoblastoma (Rb) gene in ovarian granulosa cells leads to premature ovarian failure. *Mol Endocrinol* **22**, 2141-2161 (2008).
109. Yang, Q.E., Nagaoka, S.I., Gwost, I., Hunt, P.A. & Oatley, J.M. Inactivation of Retinoblastoma Protein (Rb1) in the Oocyte: Evidence That Dysregulated Follicle Growth Drives Ovarian Teratoma Formation in Mice. *PLoS genetics* **11**, e1005355 (2015).
110. Cirulli, E.T. & Goldstein, D.B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature reviews. Genetics* **11**, 415-425 (2010).
111. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 9546-9551 (2010).
112. Kim, S. & Schliekelman, P. Prioritizing hypothesis tests for high throughput data. *Bioinformatics* **32**, 850-858 (2016).
113. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* **20**, 110-121 (2010).
114. Consortium, G.T. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204-213 (2017).
115. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
116. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics* **27**, 3216-3217 (2011).

TABLES & FIGURES

Table 1. Clinical and biochemical characteristics of study population

Subject Type	Age (years)[a]		BMI (kg/m^2)		Testosterone (ng/dL)		SHBG (nmol/L)		DHEAS (ng/mL)		Insulin (µU/mL)		LH (mIU/mL)		FSH (mIU/mL)	
Index Cases	62	29 (19-48)	62	35.6 (27.9-42.7)	62	74 (65-90)	59	56 (36-92)	61	2095 (1638-2710)	62	21 (14-29)	59	14 (7-20)	59	9 (8-11)
Fathers	62	58 (43-85)	61	28.9 (27.0-32.2)	-		-		-		49	17 (11-26)	-		-	
Mothers																
HA	6	51 (40-63)	6	34.5 (25.2-38.2)	5	51 (49-72)	5	117 (105-135)	6	2630 (1264-2694)	3	19 (15-22)	-		-	
Unaffected	2	44 (42-45)	2	24.0 (20.6-27.4)	2	32 (27-37)	2	280 (218-342)	2	1081 (901-1260)	2	12 (8-15)	2	7 (6-7)	2	8 (6-10)
Over 45 yo	54	57 (46-72)	53	28.6 (26.6-35.5)	36	26 (16-34)	25	93 (74-166)	50	645 (501-993)	28	19 (11-25)	-		-	
Sisters																
PCOS	10	29 (19-36)	10	32.3 (29.0-37.3)	10	63 (49-74)	10	55 (45-104)	10	1665 (807-2774)	10	28 (23-32)	10	10 (5-15)	10	11 (10-12)
HA	5	32 (15-40)	5	22.3 (22.3-28.5)	5	67 (54-68)	5	126 (69-177)	5	2047 (1509-3775)	5	15 (11-17)	4	4 (3-18)	4	10 (7-14)
Unaffected	57	32 (14-45)	57	25.1 (22.2-29.3)	57	30 (23-35)	55	128 (90-181)	57	1408 (1028-1814)	55	12 (9-15)	55	5 (3-10)	55	10 (7-12)
Over 45 yo	3	47 (46-49)	3	26.4 (24.5-32.1)	3	30 (20-48)	3	137 (88-150)	3	772 (765-1588)	3	13 (12-15)	-		-	
Reference Ranges																
Reproductive Aged Women	346	30 (25-35)	346	28.5 (23.0-35.4)	227	29 (21-37)	188	100 (70-144)	226	1357 (1018-1756)	185	12 (10-16)	173	4 (3-8)	173	9 (7-12)
Men	55	53 (50-57)	55	28.0 (25.9-31.0)	-		-		-		46	14 (10-17)	-		-	
Older Women	69	56 (52-60)	69	26.4 (23.2-29.6)	69	23 (15-30)	60	102 (75-156)	66	645 (487-1016)	69	13 (10-17)	48	50 (35-69)	49	80 (57-100)

Traits reported as count, median (25^{th}-75^{th} percentiles). Values reported here are unadjusted. Trait values were adjusted and normalized for variant association testing. [a]Age reported as count, median (min-max).

Table 2. Top 5 rare variant associations from quantitative trait meta-analysis

Chr	Gene	Length	Variants	Families[a]	OR[b]	p	p_{adj}
9	***DENND1A***	**550kb**	**32**	**50%**	**1.20 [0.69 – 2.14]**	**5.31×10^{-5}**	**0.019**
11	*PKNOX2*	269kb	18	39%	1.55 [0.87 – 2.94]	7.28×10^{-4}	0.27
6	*BMP6*	155kb	11	35%	1.41 [0.65 – 3.27]	3.98×10^{-3}	1.00
9	*C9orf3*	421kb	10	21%	2.79 [1.08 – 8.24]	6.14×10^{-3}	1.00
1	*PRDM2*	125kb	8	19%	3.26 [0.77 – 22.51]	6.92×10^{-3}	1.00

Rare variants exceed thresholds of predicted deleteriousness. [a]Families include all that have daughters with at least one of the rare variants. [b]Odds ratios (OR) shown with 95% confidence intervals, estimated from logistic regression of gene variant count against disease status, with equal variant weighting, adjusted for BMI, not considering relatedness.

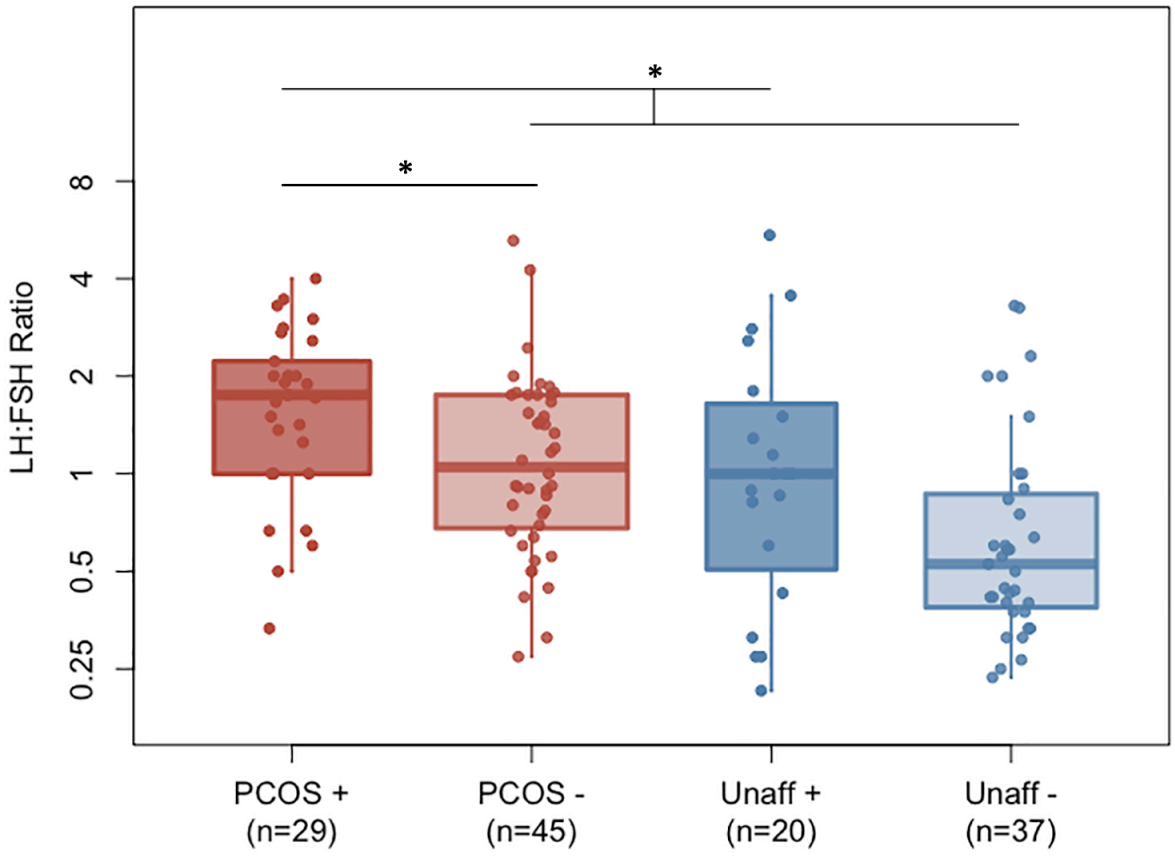

Figure 1. LH:FSH ratios in *DENND1A* variant carriers. LH:FSH ratios in DENND1A rare variant carriers (+) and non-carriers (-) in unaffected women and in women with PCOS/HA. Differences in group means were analyzed using Wilcoxon's rank sum test (* $P \leq 0.0167$).

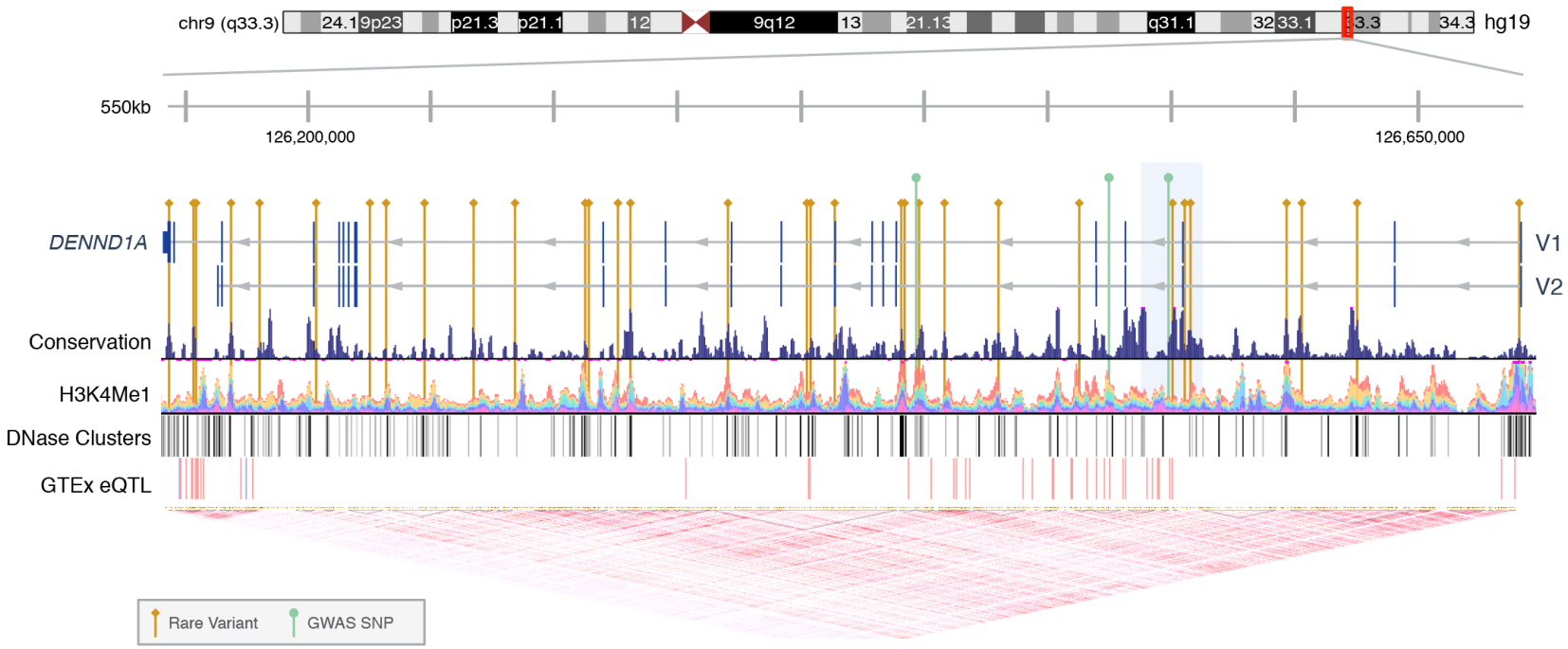

Figure 2. Rare Variants in *DENND1A*. The locations of deleterious rare variants and previously reported GWAS SNPs within the DENND1A gene, including the two primary isoforms DENND1A.V1 and DENND1A.V2. The 25kb region significantly associated with altered hormone levels is highlighted in light blue. The Conservation track was measured on multiple alignments of 100 vertebrate species by phyloP[113]. The H3K4Me1 track shows enrichment of mono-methylation of lysine 4 of the H3 histone protein, which is associated with enhancers and DNA regions downstream of transcription starts, as determined by ChIP-seq assay and layered by different cell types[53]. The DNase Clusters track shows regions of DNase hypersensitivity, an indicator of regulatory activity, with darkness proportional to maximum signal strength[53]. The GTEx eQTL track displays gene expression quantitative trait loci for DENND1A, as identified from GTEx RNA-seq and genotype data, with red and blue indicating positive and negative effects on gene expression, respectively[114]. The linkage disequilibrium heatmap was generated using Phase1 CEU data from the 1000 Genomes Project[115].

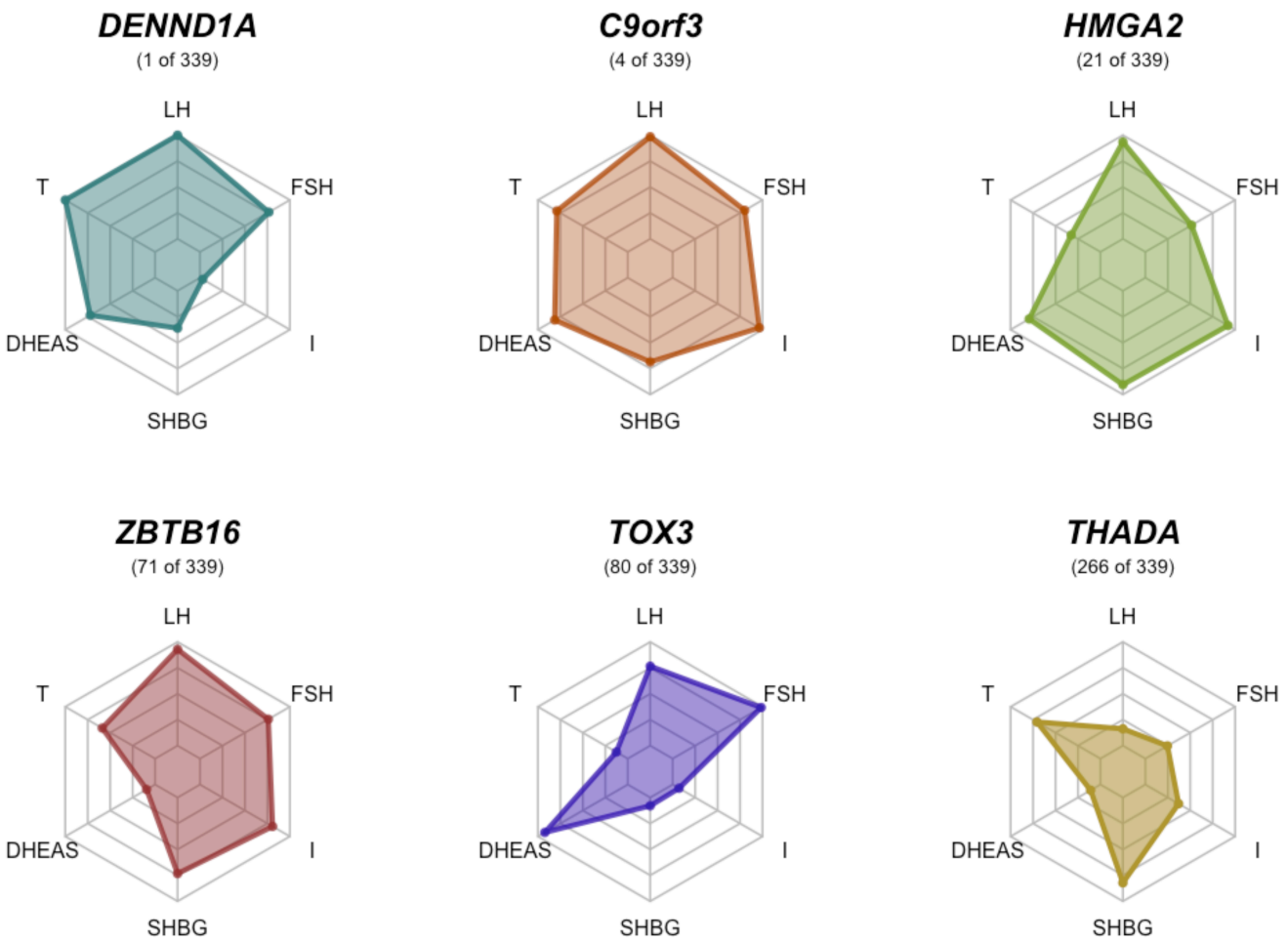

Figure 3. Trait associations for PCOS GWAS genes. The relative quantitative trait associations are shown for PCOS GWAS susceptibility loci included in meta-analysis results, with meta-analysis association ranking.

Table 3. Deleterious, rare variants in DENND1A

Position	rs ID	Variant	Allele Frequency			CADD	LINSIGHT	TF – (DISRUPTED)	RBP – (DISRUPTED)	RBP + (ENHANCED)
			Aff[1]	Unrel[b]	Pop[c]					
9:126,144,390	rs189947178	G → T	0.018	0.020	0.003	11.91	N/A	N/A	-	-
9:126,154,100	rs147370674	C → T	0.006	0.004	0.004	21.10	0.97	-	*XPO5*	-
9:126,154,582	rs529224231	T → G	0.006	0.004	0.000	21.30	0.97	*MAFK, MAFB, NRL*	-	*IFIH1, XPO5*
9:126,169,258	rs561100869	C → T	0.000	0.004	0.000	16.94	0.97	*IRF1*	*NOVA1*	-
9:126,180,758	rs750425892	G → C	0.000	0.004	0.000	19.31	0.96	*STAT6, MZF1, SPI1*	*PCBP2*	-
9:126,203,546	-	T → C	0.006	0.004	0.000	20.50	0.93	*FOXM1*	*RBMX*	-
9:126,225,586	rs538451690	T → A	0.006	0.004	0.002	16.07	0.94	-	-	*RC3H1, XPO5*
9:126,231,902	-	C → T	0.006	0.004	0.000	20.20	0.97	-	-	-
9:126,247,645	rs543947590	C → A	0.006	0.004	0.003	21.00	0.97	*STAT5A*	-	*OAS1*
9:126,267,980	rs558809288	C → T	0.006	0.004	0.000	19.97	0.84	*ARNT*	-	-
9:126,284,213	rs149244424	C → T	0.012	0.004	0.008	12.44	0.92	*ESRRA, ELF1*	-	-
9:126,312,990	rs748274474	A → G	0.006	0.004	0.003	21.10	0.96	-	-	*YBX1*
9:126,313,234	-	C → A	0.006	0.004	0.000	17.84	0.97	-	*SRSF2*	-
9:126,326,081	rs138249397	C → T	0.018	0.016	0.002	9.32	0.88	-	*OAS1*	*SRP68*
9:126,331,427	rs184609118	A → C	0.000	0.004	0.002	18.48	0.97	-	*ELAVL1, SSB*	*CELF2, NOVA1*
9:126,370,689	rs564042790	C → T	0.006	0.004	0.003	16.84	0.89	*SMAD3*	*IFIH1*	*YBX1*
9:126,402,348	rs182167487	C → T	0.000	0.004	0.000	14.18	0.93	-	-	-
9:126,404,312	-	G → A	0.006	0.004	0.000	15.72	0.91	-	*SRP54, SRP68*	*PTBP1*
9:126,414,365	rs141759269	T → G	0.000	0.004	0.001	26.00	N/A	N/A	-	-
9:126,440,857	rs147844210	T → C	0.000	0.004	0.008	21.60	0.95	*GATA6*	-	*CMTR1, FUS, SRSF3, YBX1*
9:126,441,904	rs75342773	T → C	0.006	0.004	0.008	16.64	0.95	-	*HNRNPH1-2*	*TRA2B*
9:126,447,680	rs117984673	T → A	0.018	0.016	0.008	17.70	0.94	*SOX10, SOX17*	*NOVA2*	-
9:126,458,124	rs112188193	G → C	0.012	0.008	0.013	14.43	0.92	*STAT4, STAT5A, TEAD1*	*RC3H1*	*NUDT21*
9:126,480,236	rs543924878	A → C	0.006	0.004	0.004	18.31	0.81	*SREBF1*	-	*HNRNPH1-3, HNRNPF, KHSRP, SRP14*
9:126,513,154	rs147058034	A → G	0.006	0.004	0.003	19.81	0.97	*FOXG1, FOXP3, FOXO4, FOXO6, FOXC1, FOXK1*	*CELF1, XPO5*	*NOVA1*
9:126,549,983	-	T → C	0.018	0.004	0.000	19.45	0.81	-	-	-
9:126,555,003	rs117893097	C → G	0.030	0.020	0.013	21.40	0.99	-	*HNRNPH1-3, HNRNPF, KHSRP*	-
9:126,557,679	rs552299287	A → T	0.006	0.004	0.000	18.00	0.84	*STAT5A, STAT6*	*CELF1-2, RC3H1, XPO5*	-
9:126,597,096	-	T → C	0.006	0.004	0.000	16.64	0.97	*SOX5*	*HNRNPA0, HNRNPA1, HNRNPD, ELAVL1, ZFP36*	*XPO5*
9:126,603,402	rs78012023	TTA → ATG	0.018	0.012	0.000	8.99	0.91	-	-	-
9:126,625,643	rs116887221	C → A	0.006	0.004	0.008	16.24	0.8	-	*HNRNPH1-3, HNRNPF, KHSRP*	-
9:126,691,321	rs79740971	G → A	0.006	0.008	0.010	13.52	0.84	*ESRRA*	-	-

Positions correspond to GRCh37. Variant queries were facilitated by Kaviar[116]. [a]Affected cohort allele frequencies represent proportion of alleles with variant in PCOS/HA subjects. [b]Unrelated cohort allele frequencies represent proportion of variant alleles in parents. [c] Population allele frequencies correspond to Wellderly cohort. TF- shows transcription factors for which binding is predicted to be negatively impacted. RBP- and RBP+ show RNA-binding proteins for which binding is predicted to be negatively impacted or enhanced, respectively.